|  |
| --- |
| Real Estate Analysis Project |
| The Denver Real Estate Market 2020: What&#39;s Changed Since 2019 | West Line  Village |
| September 12th, 2021  Syracuse University  IST 652  Noah Laraway |

Table of Contents

* Report Overview (page 1)
* Differences from Previous Homework (page 1)
* Data and its sources (page 1)
* Data Pre-Processing (page 2)
* Methods of Analysis (page 2)
* Overall Description of the Program (page 3)
* Analysis with Answers to Questions (page 14)
* Conclusion and Next Steps (page 15)

Report Overview

This report supports the analysis for the final project. The project is an analysis of real estate sales for Denver, CO between 2010-2021YTD and for the US from 1999-2020. It was done utilizing a csv file and an Excel file that were both read into Python using a Jupyter Notebook. The first portion of the project was to take the raw data and inspect, clean and adjust data types while converting it to data frames using Pandas. After that the data was merged and group by and aggregate functions were used to prepare for analysis. The analysis was then performed to answer questions about the data. Plots and charts were also created to visualize the data. Finally, the visualizations and resulting data frames were then exported as image files.

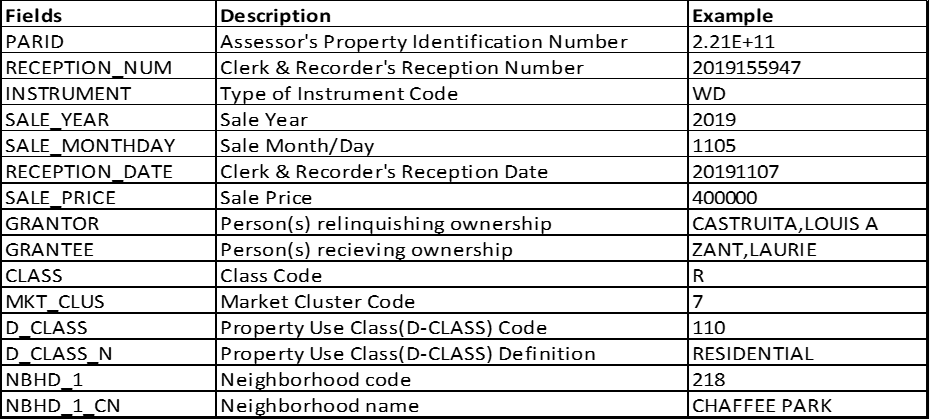
Differences from Previous Homework

* Some of the data was used for HW2, but the final project greatly expands upon what was performed before.
* This project uses both the US and Denver data sets compared to the previous assignment that only used the Denver data.
* The final project answers more questions, has more visualizations and creates more data frames than the previous assignment.

Data and its Sources

The data sets used for this project consist of 2 files:

* A CSV file from [Denver Open Data Catalog](https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-real-property-sales-and-transfers) containing property sales for the city of Denver, CO from 2010-2021YTD. The file has over 321k rows and 15 columns. Within the file there is information on sale price, property type, neighborhood and date sold.



* The second file is an Excel file from [US Census Home Sales](https://www.census.gov/construction/chars/) containing information on median and average home sales in the US from 1999-2020. The spreadsheet used for this project is the SoldMedAvgPriceDet sheet which contains 30 rows and 11 columns. Within the file there is information on year sold and the median and average sale price for home sales in the total US, Northeast, Midwest, South and West.



Data Pre-Processing

The Denver, CO real estate sales data csv was pre-processed before importing into python. The preprocessing removed columns that would not be used in the project to reduce coding that needed to be done in Python. Eight of the columns were kept in the file included: PARID (assessor’s property ID number) , Sale Year, Sale MonthDay, Sale Price, D-Class(property use code), D\_CLASS\_N(property definition), Neighborhood Code and Neighborhood Name. The updated file can be found here: [Denver\_Property\_Sales\_Updated](https://drive.google.com/drive/folders/1larh9R2MIsYZA4COTXah1JU0uTx76bDu?usp=sharing). The US sales data Excel file was not pre-processed.

Methods of Analysis

The data was cleaned, trimmed and transformed each time a question was being answered to only look at the data pertaining to that question. New data frames were created along with visualizations to support the findings.

Question #1: How have home prices in Denver and the US changed over time (2010-2020)?

* Unit of Analysis: Year (2010-2020)
* Fields Used: Year, Median Sale Price
* Resulting output is a data frame and line plot showing the change in median sales price in dollars for single family detached homes for Denver, US, Northeast, Midwest, South and Northeast. An additional analysis was performed on the percent change over this time period as well.

Question #2: What were the top 10 neighborhoods in Denver by change in sales price over time (2010-2021)?

* Unit of Analysis: Neighborhood
* Fields Used: Neighborhood, Median Sales Price
* Resulting output is a data frame and bar plot showing the median sales price in dollars for the top 10 neighborhoods in Denver over this timeframe. An additional analysis was performed on the percent change over this time period as well.

Question #3: What were the bottom 10 neighborhoods in Denver by change in sales price over time (2010-2021)?

* Unit of Analysis: Neighborhood
* Fields Used: Neighborhood, Median Sales Price
* Resulting output is a data frame and bar plot showing the median sales price in dollars for the bottom 10 neighborhoods in Denver over this timeframe. An additional analysis was performed on the percent change over this time period as well.

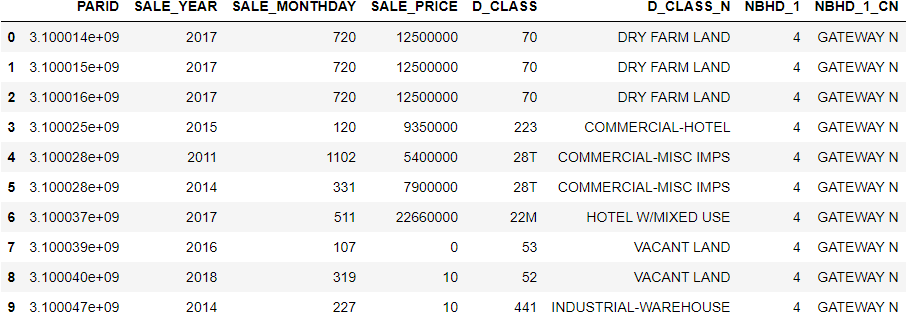
Question #4: Can we predict future sales prices for the US based of prior data (1999-2020)?

* Unit of Analysis: Year (1999-2020)
* Fields Used: Year, Median Sale Price, Forecasted Sale Price
* Resulting output is a line chart showing the moving average of sale prices in the US for 2-, 3- and 5-year moving averages. An additional plot was created using the 2-year moving average to predict prices over the next 5 years.

Overall Description of the Program

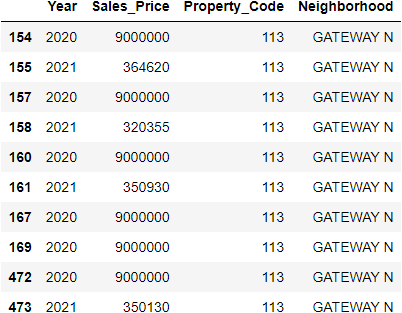
The initial portion of the program imports, inspects and cleans the data. For importing the data files pandas read\_CSV and read\_excel functions were utilized. After that an inspection of the Denver real estate data was performed and it was discovered that several changes to the data frame needed to be completed.

Raw Data for the Denver Real Estate Sales:

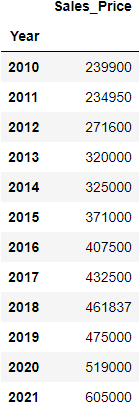
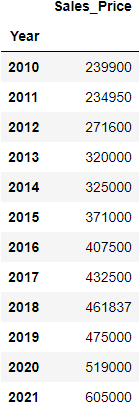


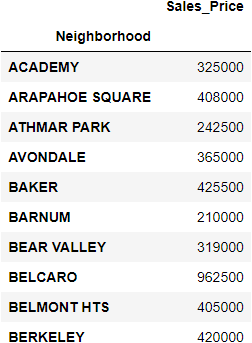
This data file contains sales for all properties including commercial and farm land. Since this project focuses on single family homes the other categories needed to be removed. This was completed by filtering down to the single-family home codes (D\_Class = 113 through 119). Also, there were quite a few sales that were $0 or $10 meaning they were most likely a quit claim, but definitely not a typical sale. These sales were filtered out by removing all sales less than $1000. It was then decided that columns for PARID, SALE\_MONTHDAY, D\_CLASS\_N & NBHD\_1 were unnecessary and removed. The remaining column headers were changed for easier reading ability.

Resulting Cleaned Data Frame:

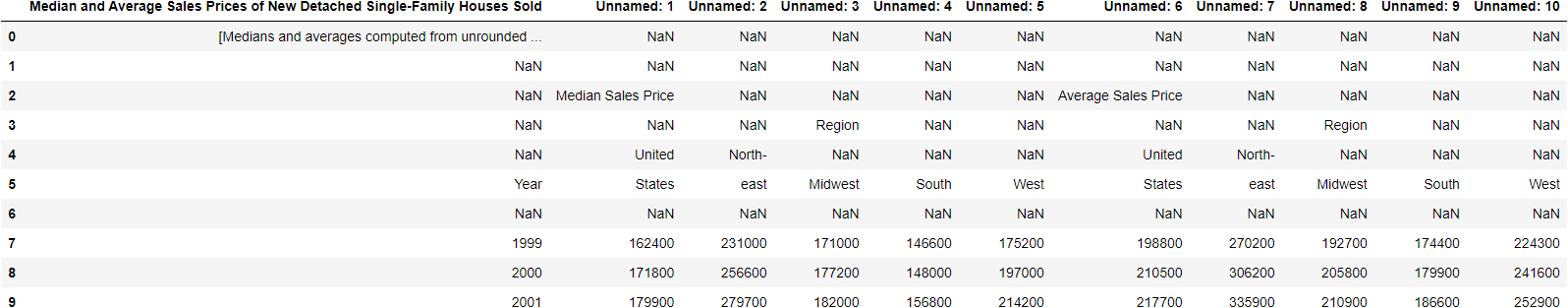


From there two additional data frames were needed from the Denver data frame that would be used for further analysis. The first data frame was for median sales price by year using a group by function. The second also used a group by function, but was for median sales price by neighborhood.

Data Frame for Median Sales Price by Year: Data Frame for Median Sales Price by Neighborhood:

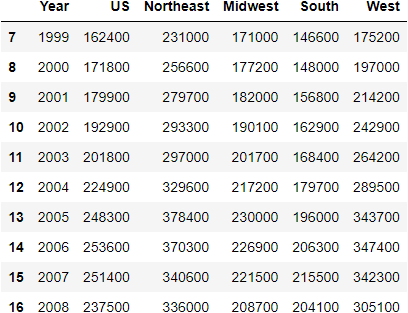


The US real estate data frame was then inspected and it was discovered that it too had to be cleaned.

Raw Data for the US Real Estate Sales:  


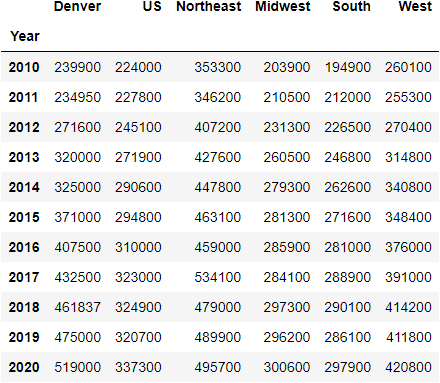
The excel file that was brought in contained unnecessary rows and columns that were removed using the drop function in Python. The columns were then renamed for better reading ability and the datatypes were changed from object to integer.

Resulting Cleaned US Data Frame:



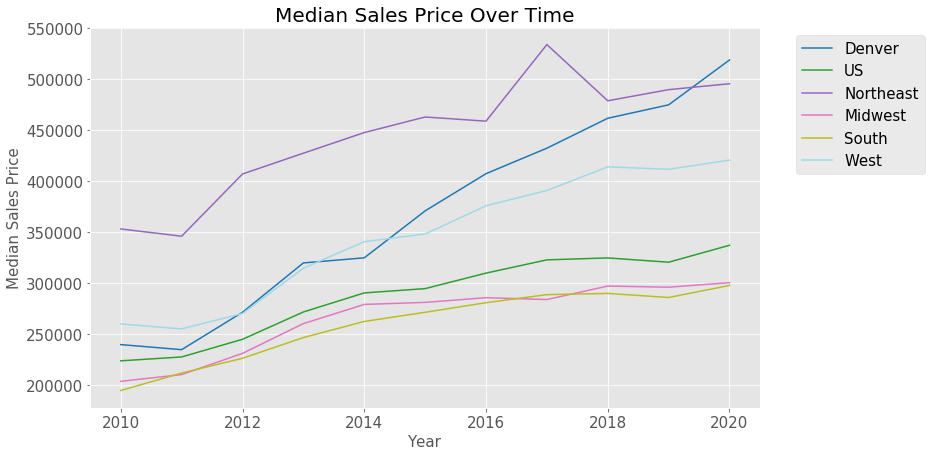
At that point both of the data sets were ready for the next step of performing the analysis to answer the questions. To answer the first question on how home prices have changed in Denver and the US over time the US cleaned data frame and the Denver data frame for median sales by year were merged together. For US data the timeframe was from 1999-2020, but for the Denver data it was 2010-2021YTD. The merged data frame only contained the overlapping years of 2010-2020.

Merged data frame:



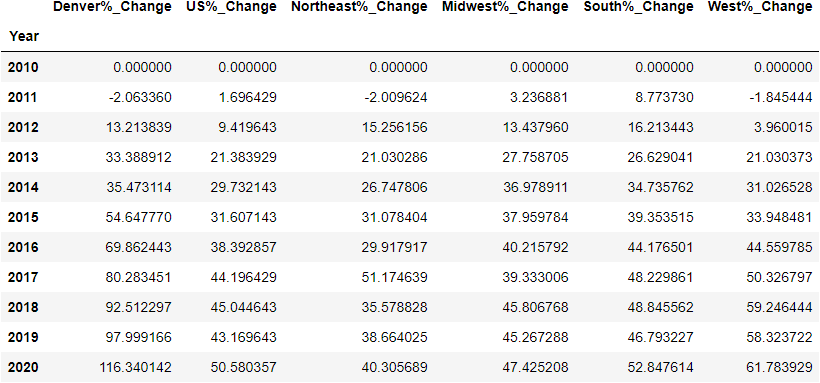
A line plot was also created using ggplot for each of the 6 locations over the time period with median sales price on the y-axis and year on the x-axis.

Line plot for Median Sales Price over time:

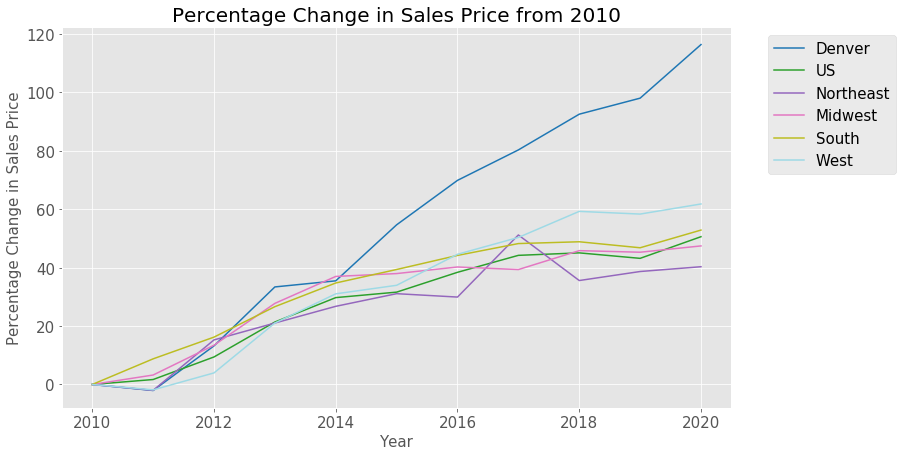


Another data frame and plot were created for the percentage change in median sales price. The data frame was made by adding percentage change columns to the merged data frame using iloc in Python and then dropping the original median sales price columns.

Data frame for percentage change in median sales price:

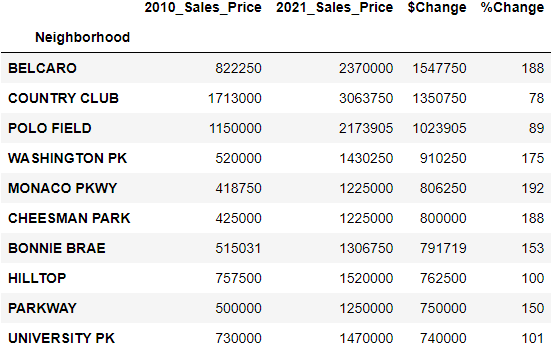


Line plot for percentage change of median sales price over time:

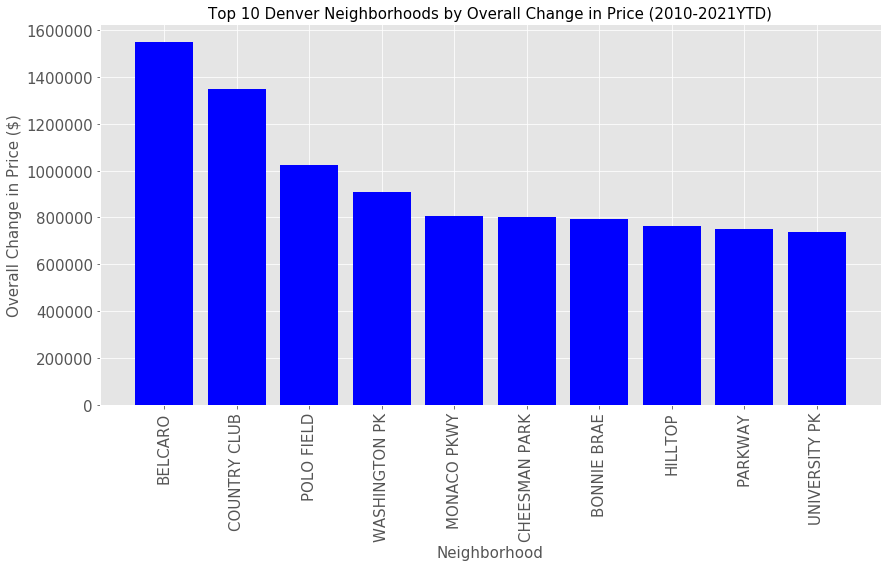


The next question to answer was what were the top 10 neighborhoods in Denver by change in sales price over the time period (2010-2021)? The Denver neighborhood data frame was used and a group by function was performed on the year and neighborhood columns and then aggregated by median sales price. This resulted in a data frame that had the median sales price for each neighborhood by year. To normalize the data, I only looked at neighborhoods that had sales in 2010 and 2021 for the comparison. A data frame for 2010 and 2021 sales by neighborhood was created. This data frame was then sorted by dollar change and filtered to the top ten neighborhoods. A bar plot was also created to visualize the data.

Data frame for top 10 neighborhoods by dollar change:

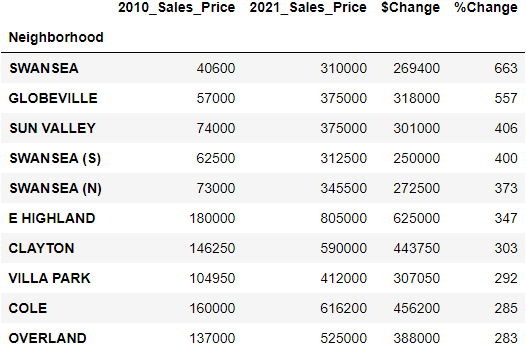


Bar plot for top 10 Denver neighborhoods by overall dollar change:

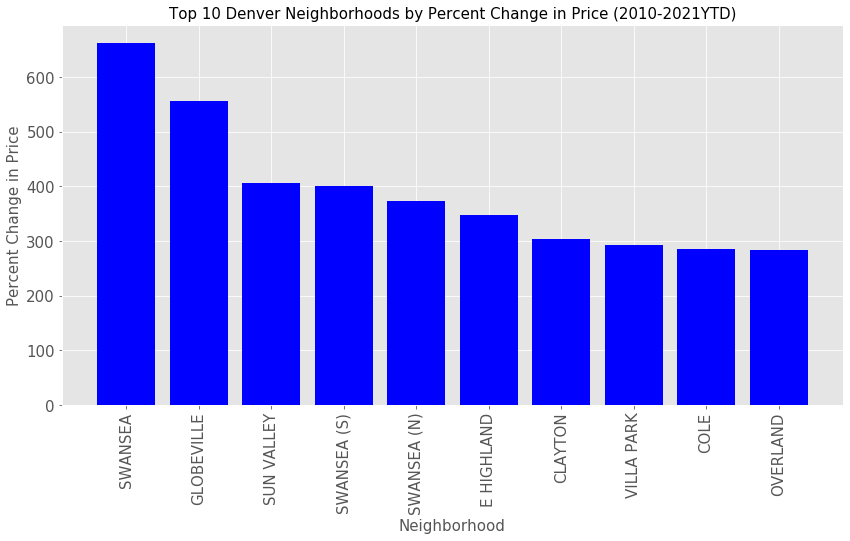


A data frame and bar plot were also created for the top 10 neighborhoods by percentage change over the time period.

Data frame for top 10 neighborhoods by percentage change:

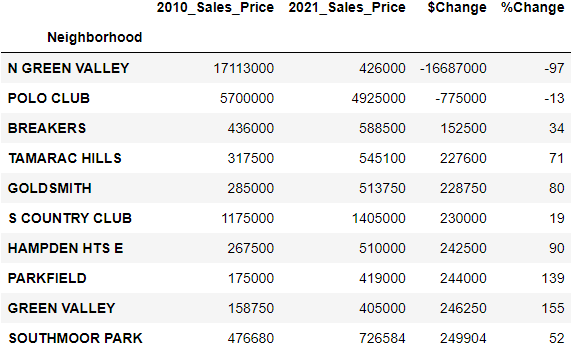


Bar plot for top 10 Denver neighborhoods by percentage change:



The next question to answer was what were the bottom 10 neighborhoods in Denver by change in sales price over the time period (2010-2021)? The same data frame for the top 10 neighborhoods was used, but with ascending being equal to true to get the lowest median sales increase by dollars.

Data frame for bottom 10 neighborhoods by dollar change:

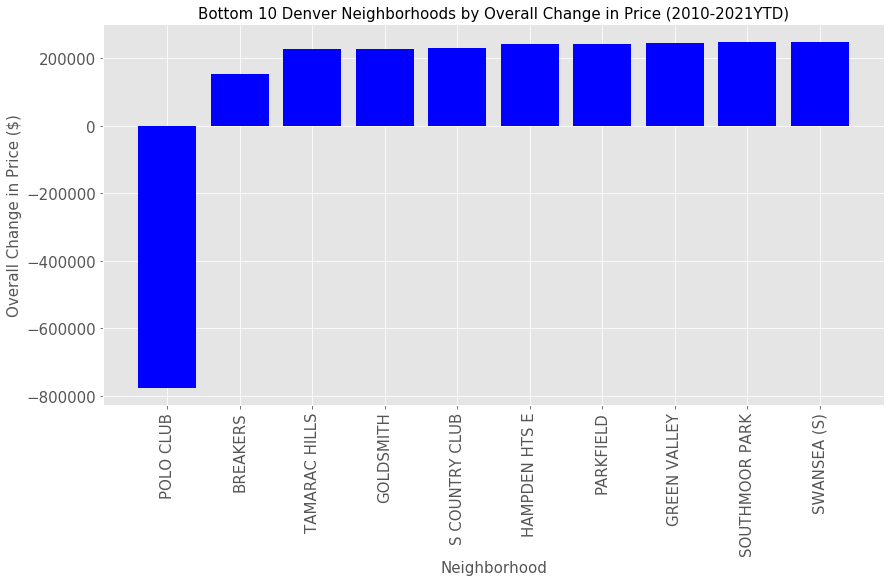


It is clear that there is an issue with the neighborhood N Green Valley. It says the median sales price in 2010 was over $17MM, but in 2021 it was only $426K. This neighborhood is relatively new and was in the early stages of development in 2010. Most likely the 2010 sales were due to home builders either buying multiple lots in one sale or buying large amounts of acreage in one sale. For this project it was decided to remove this neighborhood from the analysis. The N Green Valley neighborhood was filtered out and an updated data frame and bar plot for bottom 10 neighborhoods by overall sales price was created.

Updated data frame for bottom 10 neighborhoods by overall change:

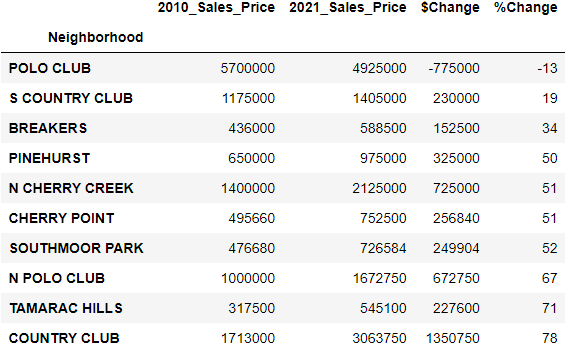


Bar plot for bottom 10 Denver neighborhoods by overall change:

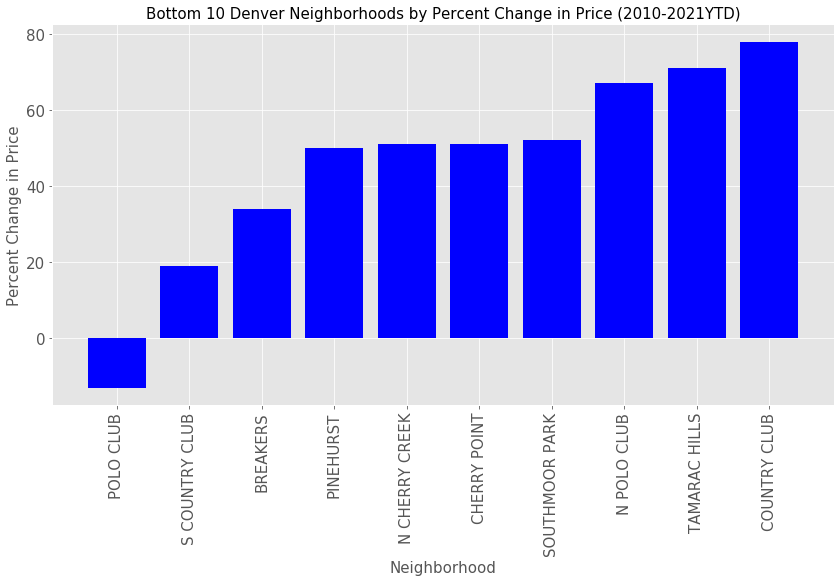


A data frame and bar plot were also created for the bottom 10 neighborhoods by percentage change over the time period.

Data frame for bottom 10 neighborhoods by percentage change:

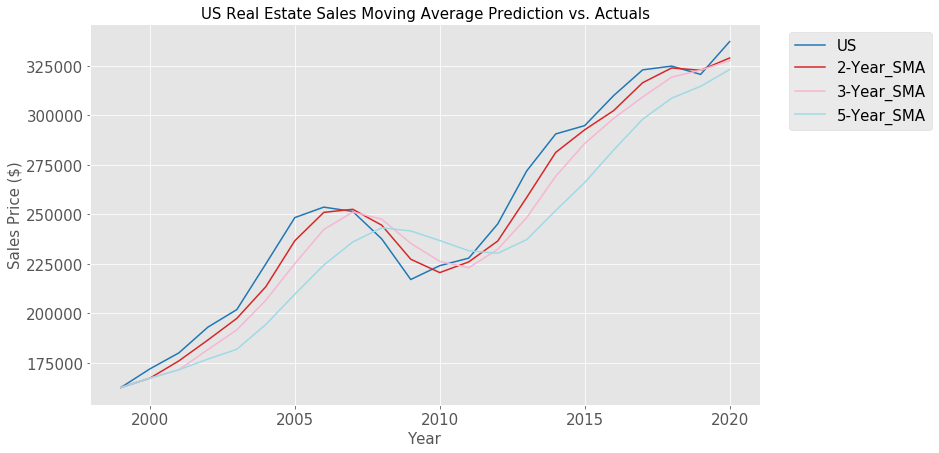


Bar plot for bottom 10 Denver Neighborhoods by percentage change:



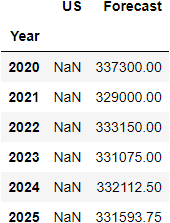
The last question to be answered was can we predict future sales prices for the US? Simple moving averages for the US were created that trailed the data by 2, 3, and 5 years. The US data frame was filtered down to just US overall median sales prices by year from 1999-2020. The moving averages were added to the data frame using the rolling function and then a line plot was created.

Line plot for US sales with moving averages:



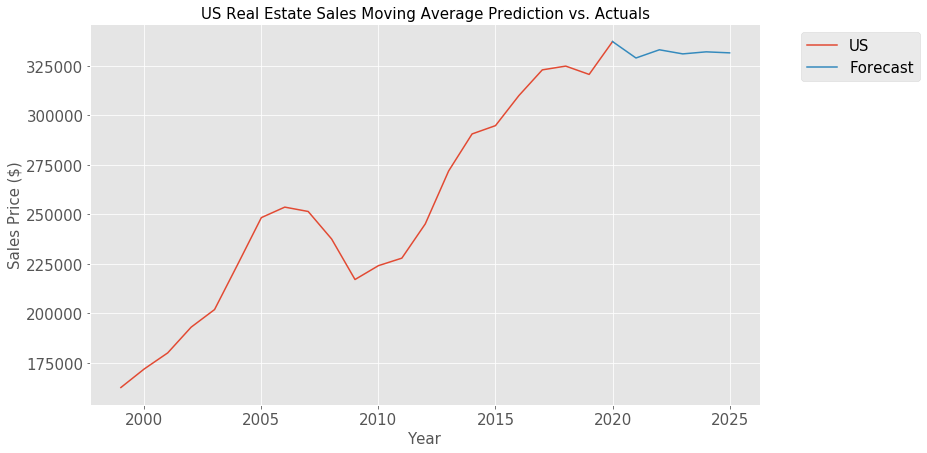
Looking at the plot the 2-year moving average performed the best and was used for the forecasting through 2025. The 2-year moving average forecast was performed using a for loop and put into lists and then into a data frame. The forecasted data frame was then added to the US moving average data frame and the moving averages were removed to only show the US actuals and the forecast.

Forecasted data frame through 2025:



A line plot was created using ggplot showing the US actual median sales data from 1999-2020 and the forecasted data.

Line plot with US actuals and forecast:



Analysis with Answers to Questions

Question #1: How have home prices in Denver and the US changed over time (2010-2020)?

* The median sales price for homes has gone up drastically in Denver and in all regions of the US from 2010-2020. For the US as a whole the percentage increased by 50.6% over the timeframe. In Denver the change was the most dramatic with an increase of over 116%, more than double the percent increase of the US as a whole. This is intuitive due to the rapid increase in population in Denver over this time frame and the relatively low median sales price in 2010 for Denver.

Question #2: What were the top 10 neighborhoods in Denver by change in sales price over time (2010-2021)?

* The top 10 neighborhoods by overall change in price in descending order were: Belcaro, Country Club, Polo Field, Washington Park, Monaco Parkway, Cheesman Park, Bonnie Brae, Hilltop, Parkway and University Park. All of these neighborhoods have a median sales price of over $1MM in 2021. Almost all of these neighborhoods are well known in Denver as being highly affluent and it is no surprise to see them on this list. Also, the majority of these neighborhoods are in close proximity to each other on the east side of Denver.
* The top 10 neighborhoods by percent change in descending order were: Swansea, Globeville, Sun Valley, Swansea (S), Swansea (N), E Highland, Clayton, Villa Park, Cole and Overland. These areas have been historically known in Denver as being lower socioeconomic neighborhoods, but have seen a lot of growth and gentrification in the last ten years. It does make sense to see the gentrifying neighborhoods on this list instead of the more affluent areas since they have more room for growth by percentage increase. Also, the majority of these neighborhoods are in close proximity to each other on the north side of Denver.

Question #3: What were the bottom 10 neighborhoods in Denver by change in sales price over time (2010-2021)?

* The bottom 10 neighborhoods by overall change in price were: Polo Club, Breakers, Tabarac Hills, GoldSmith, S Country Club, Hampden Hts E, Parkfield, Green Valley, Southmoor Park and Swansea (s). This list contains some highly affluent and also some lower socioeconomic neighborhoods. The affluent neighborhoods like the Polo Club and S Country Club are most likely in the bottom ten due to already being highly priced back in 2010 and did not see a large increase in value over the time frame with the Polo Club neighborhood actually decreasing in price. The lower socioeconomic neighborhoods on the list did see a large percentage increase in value over the time frame, but the overall increase in dollars was not as substantial compared to the rest of Denver.
* The bottom 10 neighborhoods by percent change were: Polo Club, S Country Club, Breakers, Pinehurst, N Cherry Creek, Cherry Point, Southmoor Park, N Polo Club, Tamarac Hills and Country Club. This list contains more affluent neighborhoods and even though some of them had a large increase by dollar amount, it was relatively low percentage wise compared to the rest of Denver

Question #4: Can we predict future sales prices for the US based on prior data (1999-2020)?

* Although not required for the project, I did want to include some time series forecasting with the data. Having not done this in Python before, it was somewhat of a steep learning curve for me. I originally wanted to use a more advanced forecasting model and tried using ARMA(Auto-Regressive Moving Average) from the statsmodel module. Unfortunately, I was unable to get a good fit with ARMA. Ultimately, I decided to remove it from the program and use the simple moving average instead.
* I plotted a 2, 3 and 5-year moving average against the US median sales price and it was clear that the 2-year moving average trended the best with the actual data. Forecasting out to 2025 with the 2-year moving average predicts that prices will start to level off soon and potentially decrease in the next 5 years.

Overall Conclusions and Next Steps

With the drastic change in recent real estate prices throughout the US and in Denver I thought this would be an interesting topic to explore. It was intriguing to dig through the data and see how prices have changed by region and neighborhood and how they compare to one another. The analysis of the data confirmed some initial assumptions, but there were some findings that were unexpected. When performing the forecast of the time-series I found it challenging, but it did give me an idea of how it can be used with Python.

It was eye opening to see how many real estate sales there were in Denver over this time frame with over 321k rows of data. It was also insightful how much of the data is not needed and how important it is to spend extra time up front deciding which data to use and how to filter it down. Even after inspecting and cleaning the data there was still issues found during the analysis like the Denver neighborhood of N Green Valley having a median sales price of over $17 million in 2010. Understanding the data and being attentive to possible data issues is critical for performing a high-quality analysis.

The analysis portion of the project and the visuals created gave a good picture of recent real estate trends. Prices have increased in all regions of the US and Denver has increased at an even higher rate. Due to the population boom in Denver over the last 10 years this was expected, but seeing a 116% increase in ten years was a lot more than I would have predicted.

The analysis on the Denver neighborhoods was interesting as well. Some of the most affluent neighborhoods had high increases in overall value, but not in percentage change. This was more than likely due to the affluent neighborhoods already having high home values in 2010 and not being gentrified since then. The top neighborhoods by percentage change ended up being all lower socioeconomic areas. This shows how some of the poorer neighborhoods in Denver have changed recently due to gentrification.

When performing the time-series forecasting I found it to be more difficult than I originally thought it would be. I wanted to use a more sophisticated model, but had a tough time getting a good fit with ARMA. I decided on the simple moving average model due to its simplicity. Although other models could have potentially forecasted more accurately, I was able to get some predictions with the moving average model and get it plotted. Even though I had some difficulty with the forecasting portion of the project, it was a valuable experience and gave me some introduction to doing time-series predictions in Python.

Going forward I plan to continue to add additional years of data to the project to see how real estate prices evolve over time. It would also be interesting to add data from other cities and see how it compares to Denver and the US. Also, getting more experience with forecasting and utilizing it in Python will be highly beneficial to making time-series predictions. Overall, the project was extremely useful for me in gaining experience utilizing Python for data cleaning, transforming and performing analysis.